Inferring marketing channel relevance in the customer journey to online purchase

David A. Wooff[a*] and Jillian M. Anderson[b]

[a] *Durham University, Department of Mathematical Sciences, Stockton Road, Durham DH1 3LE, UK*
[b] *Summit Media Ltd, Albion Mills, Albion Lane, Willerby, Hull, HU10 6DN, UK*

July 7th 2013

Abstract

In this paper we address the problem of inferring marketing channel importance for the customer journey to online purchase, using sequential data analysis ideas. We suggest a method for inferring the relative value of channels using historical data. We propose metrics for source, intermediary, and destination channels based on two- and three-step transitions in fragments of the customer journey. We comment on the difficulties of formal hypothesis testing. We illustrate the ideas and computations using data from a major UK online retailer.
Keywords: Sequential analysis; Channel relevance; Online marketing; Path to conversion; Clickstream; Digital marketing; E-commerce.

1 Introduction

This paper concerns analysis of the sequential data representing routes to online purchase – known as conversion – by customers at a retail internet site. We describe the background in a companion paper Wooff and Anderson [2013]. The process is also fully described in Abhishek et al. [2012] and Xu et al. [2012]. Briefly, a customer may visit several websites, including multiple visits to the final retail site, before converting. These visits are captured using cookies stored on consumers' computers. The various sites visited are classified as marketing channels of different kinds, for example social media and direct email. Digital advertising spend in the UK alone amounted to £5416 million in 2012, with annual growth of around 13% [Internet Advertising Bureau UK, 2013]. As such, there is enormous interest in determining which channels are relevant to the final purchase. One reason is because some part of the sales revenue is attributed to channels in the customer journey, for example shopping comparison sites may be rewarded for funnelling customer traffic through to the final retail site. This is the problem known as weighted attribution which is the focus of our companion paper Wooff and Anderson [2013]. A second reason is that the different marketing channels might be stages in, or different aspects of, an advertising campaign, and where it is desired to measure the value of each aspect in contributing to the final purchase decision.

There are a number of algorithm-based methods which use converting and non-converting journeys in order to determine the probability of each channel leading to a conversion; for example, see Shao and Li [2011]. Abhishek et al. [2012] view the journey as a funnelling process whereby customers are influenced by typically narrower funnels at each step by the marketing material. They address the likelihood to convert at each stage and then present an attribution scheme based on the increment that each step has on the consumer's probability to convert. They take a large set of data from an online campaign for a large car manufacturer and construct a hidden Markov model to relate advertising stages to conversion behaviour. This is useful for tightly-defined advertising campaigns. Xu et al. [2012] view the journey as a Markov process with a special structure – mutually exciting point processes – and so fit models which result in a measure of each channel's value as well as allowing prediction of conversion rate. Both these methods require

*Corresponding author. Email: d.a.wooff@durham.ac.uk

conversion and non-conversion histories which arise because each advertising stimulus can be assessed as leading definitely to a conversion or, in a time-censored sense, to a non-conversion.

Our interest is in data which is less clean. We consider all journeys which end in a conversion for a particular retailer, from whatever source. we do not consider non-converting journeys as we have no data concerning them, as is standard in data of this kind. We cannot analyse journeys which end at a different retailer. We may analyse fragments of journeys in which a customer visits a particular retailer, but does not make purchase, but doing so requires many quite deep assumptions which reflect factors concerning a particular retailers position within the marketplace. In other words, we may analyse only what we have observed and there is no element of experimental design involved - for that, the methods described in Abhishek et al. [2012], Xu et al. [2012] are more appropriate.

For an introduction to statistical methods to discover statistically surprising patterns in sequences see for example Agrawal and Srikant [1995], Zaki [2000a,b], Wang and Yang [2005]; however this is not central to our problem of inferring channel relevance. The main focus in Agrawal and Srikant [1995] is to find customer journeys which have a specified minimum level of support, each such journey being classified as a sequential pattern; a subsidiary focus is on which items are purchased as part of the same journey. See Hahsler et al. [2005] for a more recent discussion of mining of association rules and a computer package providing tools. The problem of predicting the next step in a journey conditional on the observed history is also much studied, but is not relevant here. For predicting from a clickstream history, see for example Gunduz and Ozsu [2003], Gunduz-Oguducu and Ozsu [2006]. There is also a literature on exploring web navigation behaviour; these tend to focus on website analytics. Berendt and Spiliopoulou [2000], for example, use knowledge of local web infrastructure with sequential pattern analysis to assess site design. Other researchers have used Markov and Hidden Markov models to construct predictions for customer browsing behaviour; see Jamalzadeh [2012] for an overview.

2 Principles and notation

We will employ a notation based on that of Agrawal and Srikant [1995]. Our concern is with journeys which interact with a fixed number, n, of nodes $X_1, X_2, \ldots, X_n$ in some order. In common with the digital marketing community, we call these interactions visits. The journeys may contain loops, repeated fragments, and so forth. There may or may not be single-click journeys. We describe cleaning of the data in Wooff and Anderson [2013], noting that journeys are typically left-censored to the most recent S steps, so that S is the maximum sequence length. Let $A \to B$ mean the direct transition from node A to node B. Let $A \Rightarrow B$ mean any one-step or two-step transition from A to B. The notation $\bar{B}$ means any node except node B. Let $N\{ij\}$ be the number of times the direct transition $X_i \to X_j$ occurs. We extend the notation to longer sequences, so that $N\{ijk\}$ is the number of times the subsequence $X_i \to X_j \to X_k$ appears.

2.1 Cyclic sequences

Ideally we want to deal with uniquely classified nodes, for example a unique landing page within a retail website. In this situation it makes sense to treat a sequence $(A \to A \to B)$ as equivalent to the sequence $A \to B$, such that the sequence then contains no immediate loops, and we do not distinguish between one interaction and more than one interaction with the node. This principal seems to extend naturally to subsequences. That is, $(A \to B \to A \to B)$ might be considered equivalent to $(A \to B)$. Ultimately this is the restriction that the sequence not be cyclic. However, there are difficulties in working with this interpretation. First, checking for cyclicity is non trivial [Wang and Yang, 2005]. Secondly, if the journey is actually cyclic, we need to decide which part of the journey to disregard.

In determining channel relevance, the possible nodes in many examples happen to be crude bins representing channel type rather than a granular classification. Therefore it is perfectly feasible to observe a journey such as $A \to A$, for example from one shopping comparison site to another. Thus in the remainder of this account we make no sequence restrictions and allow sequences to be cyclic.

3 Inferring node value using sequential analysis

The key concept is whether a node contributes to a journey which ends in a sale. Clearly, the final nodes in the journey are important, but time-weighted attribution of revenue as discussed in Wooff and Anderson [2013] will emphasize these anyway. Therefore in what follows, we will derive relevance of node independently of early or late position in the journey. The proportion of nodes visited across all journeys offers a simple measure of relevance. However, the key is to measure the importance of a node in terms of moving from one to another. Thus, we need to focus on the probabilities of transition. Thus, suppose the customer journey includes the sequence $A \to B \to C$. The questions to answer are: how relevant is the intermediary node B, and would the customer have reached C from A regardless? Ideally we would like to represent customer journeys using probabilistic networks such as Bayesian belief networks; however these are inadequate for the task, partly because they are directed networks and partly because their inherent Markov properties cannot handle multinode histories.

We must take into account at least three-step transitions. This is already challenging; dealing with all possible four-step transitions, where we would have to consider all possible intermediary pairs of nodes, is daunting. Thus we restrict attention to two steps and three steps. We will ignore whether fragments of a journey occur early or late. We will remove single-step journeys from consideration as these are not informative for transitions. For each sequence we now construct two-step and three-step fragments as follows. Take as an example the sequence:

$$A \to B \to C \to B \to E \to D.$$

These contain these two-step fragments:

$$A \to B, \;\; B \to C, \;\; C \to B, \;\; B \to E, \;\; E \to D,$$

and the three-step fragments:

$$A \to B \to C, \;\; B \to C \to B, \;\; C \to B \to E, \;\; B \to E \to D.$$

Thus a journey with length s contains $s-1$ two-step fragments and $s-2$ three-step fragments. Clearly by breaking down journeys into fragments we are losing much information, particularly about more complicated journeys.

3.1 A metric for intermediary node value

We now propose a metric for channel relevance. A natural metric for the relevance of a node B in journeys from A to C is the proportion of such journeys which pass through B, which we estimate by the observed proportion:

$$\Lambda_{ABC} = \frac{N\{ABC\}}{N\{AC\} + N\{ABC\} + N\{A\bar{B}C\}}.$$

This is the observed conditional probability that any two- or three-step journey from A to C passes through B, $P(A \to B \to C | A \Rightarrow C)$. If this value is small, it suggests that B is not an important way of reaching C from A. If this value is large, it suggests that B is an important intermediary. More formally, for a (source, intermediary, destination) triple this metric is:

$$\Lambda_{ijk} = \frac{N\{ijk\}}{N\{ik\} + \sum_{j=1}^{n} N\{ijk\}}, \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, n, \;\; k = 1, \ldots, n.$$

A general measure of the value of node X_j is then given by averaging over all source and destination nodes:

$$\lambda_j = \sum_{i=1}^{n} \sum_{k=1}^{n} \Lambda_{ijk}, \quad j = 1, \ldots, n. \tag{1}$$

Note that these measures do not sum to unity:

$$\sum_{j=1}^{n} \lambda_j = \sum_{i=1}^{n} \sum_{k=1}^{n} \frac{v_{ik}}{1 + v_{ik}} \leq 1, \;\; v_{ik} = \frac{\sum_{j=1}^{n} N\{ijk\}}{N\{ik\}}, \tag{2}$$

where v_{ik} is the ratio of indirect to direct transitions for node pair (i,k). This sum depends on the total number of direct two-step transitions and the total number of exactly three step transitions for each node pair. Thus, a normalized metric is given by

$$\tilde{\lambda}_j = \lambda_j / \sum_{j=1}^{n} \lambda_j. \tag{3}$$

As a simple average, (3) does not take into account the volumes of journeys between pairs. As such, a refinement is to weight according to volume. Typically we deem the destination node to be more relevant than the source node so that it can be appropriate to weight according to the volume of destination nodes. It is trivial to weight according to other choices of volume. Let z_k be the number of two-step journeys which end at node k, and let z_0 be their sum, i.e. the total number of two-step journeys. That is,

$$z_k = \sum_{i=1}^{n} N\{ik\}; \quad z_0 = \sum_{k=1}^{n} z_k.$$

Then $\tilde{z}_k = z_k/z_0$ is the proportion of two-step journeys which end at node X_k, with $\sum \tilde{z}_k = 1$. This gives a relative measure of the volume of destination node X_k. Now a plausible measure of the value of intermediary node X_j is

$$r_j = \sum_{i=1}^{n}\sum_{k=1}^{n} \tilde{z}_k \Lambda_{ijk}, \quad \tilde{r}_j = \frac{r_j}{\sum_{j=1}^{n} r_j}, \tag{4}$$

where the latter is normalized. If we also wanted to take into account the value of the source node X_i via some weight $\tilde{y}_i$ with $\sum \tilde{y}_i = 1$, then (4) is easily extended to

$$r_j^* = \sum_{i=1}^{n}\sum_{k=1}^{n} \tilde{y}_i \tilde{z}_k \Lambda_{ijk}, \quad \tilde{r}_j^* = \frac{r_j^*}{\sum_{j=1}^{n} r_j^*}. \tag{5}$$

In the main example of Wooff and Anderson [2013], we use (4), so that the normed value $\tilde{r}_j$ is our principal metric for determining the relevance of intermediary node j.

3.2 Metrics for the journey relevances of initiating and terminating nodes

We may develop similar metrics to value different features of a journey. The two most useful are as follows. The proportion of journeys from B to C which are preceded by A is estimated by their observed proportion:

$$\Phi_{ABC} = \frac{N\{ABC\}}{N\{BC\}}.$$

If this value is small, it suggests that A is not an important way of starting $B \to C$ journeys. Note that this metric ignores direct AC transitions, and so can't be used as a measure of the importance of A in the journey to C alone. The proportion of journeys from A to B which continue on to C is estimated by:

$$\Psi_{ABC} = \frac{N\{ABC\}}{N\{AB\}}.$$

A high proportion suggests that most customers did not find B a suitable place to stop. A high proportion could also imply that B is a natural way of getting to C. For each of these metrics, we may weight and normalize according to volume as desired.

3.3 A note on tests of uniformity

It is possible but challenging to develop statistical tests of uniformity for such metrics. As such, we will not carry out these tests, but comment on some of the issues. Conditional on ending at X_k and starting at X_i we have $N\{ik\} + N\{ijk\} + N\{i\bar{j}k\}$ possible journeys of which $N\{ijk\}$ went through X_j. This is like imagining that someone at X_i wants to get to X_k but isn't sure how to get there. We might then assume that the

total number who end up at X_k via X_j is binomial $b(N, p)$ with parameters $N = N\{ik\} + N\{ijk\} + N\{i\bar{j}k\}$ and unknown probability p estimated as λ_{ijk}. This leads naturally to a standard error for the estimate as

$$s_{ijk} = \sqrt{\frac{\lambda_{ijk}(1 - \lambda_{ijk})}{N\{ik\} + N\{ijk\} + N\{i\bar{j}k\}}}. \tag{6}$$

We can do this for each node separately, and for all n^2 combinations of beginning and ending nodes. However this ignores a degree of correlation between the measures. Instead, if we can make the same assumption about N being fixed, we can treat the outcomes as multinomial for a fixed starting and ending pair. The outcomes then are all routes which pass through an intervening node plus the direct route probability. Thus, for any pair of nodes X_i, X_k, let $N = N\{ik\} + \sum_j N\{ijk\}$. This is the total number of routes from X_i to X_k either direct or via one intervening node. Before we observe the data, N is a random variable. As such, the following results are conditional on N being fixed at what was observed. Now let p_0 be the probability that a route starting at X_i and determined to get to X_k goes directly, and p_j the probability that such a route passes through node X_j. These probabilities may be routinely estimated using the multinomial distribution. A test of uniformity is easily provided by a Chi-squared test. However, the test:

$$H_0 : p_0 = p_1 = \ldots = p_n$$

versus the alternative that at least one p_i differs is not so interesting. This is because we would generally expect a much higher probability p_0 for the direct transition. Therefore, attention should focus on the hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_n,$$

i.e. that the indirect transition probabilities are all equal. A second problem is the number of tests we would need to carry out. If there are n nodes, we would need to carry out n^2 hypothesis tests, which would be correlated, and then it is doubtful that we would wish to analyse the results of all of these in detail. Finally, the nature of the data implies very unbalanced sample sizes. Some of the pairs could be associated with such large volumes of data that spuriously small p-values result, whereas for others there may be no or little data. As such, an effect-size approach [Wooff and Jamalzadeh, 2013] may be more useful – this is a focus of future research.

4 Computation and illustration

INSERT Table 1 about here

As an example, we explore data from a major UK online retailer. This records 58667 journeys of which 27420 are single-click and 31247 have at least two clicks. 17841 journeys have at least three clicks. We limit to the most recent $S = 19$ steps of any journey. Each click is classified as belonging to one of nine channels as shown in Table 1. This shows that a high proportion of single-click journeys for this retailer at this time were branded natural search, coded as **NatB**.

We now take every journey and count all the pair occurrences. The counts are shown in Table 2 and plotted as a balloon plot in Figure 1, with area proportional to count. There are 83387 pairs. Again, the **NatB** node dominates, and there are several nodes which carry very little traffic. Also of interest is the conditional bivariate transition matrix plotted in Figure 2. This shows the proportion being received by each receiving node given the sending node, i.e. $P(X_i \to X_j | X_i$ is the sender). Probabilities across rows sum to one. (Interpreting columns is not sensible.) There are two obvious deductions we make from Figure 2. First, there is a high probability of clicking on the same kind of channel, i.e. $A \to A$, regardless of where you start. This is evidenced by a strong diagonal pattern. Secondly, there are high conditional probabilities of ending in nodes **Aff** and **NatB** regardless of starting node, as evidenced by high probabilities in those columns. Understanding of such patterns is obviously useful for marketing design and so forth, but is not our focus here.

INSERT Figure 1 about here

INSERT Figure 2 about here

INSERT Table 3 about here

We next address journey triples. There are $n = 9$ possible intermediary nodes for each sender and receiver. Table 3 counts the number of triples where the intermediary node is **NatB**. There were overall 94 journeys **Aff** $\rightarrow$ **NatB** $\rightarrow$ **Nat** and no journeys **List** $\rightarrow$ **NatB** $\rightarrow$ **Ban**.

We may assess whether the starting and ending nodes of two-step patterns resemble in frequency the starting and ending nodes of three-step patterns. To do this, count for each pair of nodes i, k, the number of direct transitions $N\{ik\}$ and the number of indirect transitions $\sum_j N\{ijk\}$ via any intermediary node. Table 4 shows the number of indirect transitions, and Table 5 shows the proportions v_{ik} of indirect transitions to direct transitions obtained by dividing Table 4 by Table 2. On average this ratio is 66%. We can carry out a Cochran-Mantel-Haenszel test to test for differences between these two tables. This test is strongly significant; we conclude that the two tables have different patterns. However, the statistical significance is partly the result of very large sample sizes. Indirect transitions to **Aff**, **Comp**, **Un** tend to occur relatively less than average, and indirect transitions to **Ban**, **List**, **PPC** tend to occur relatively more than average. Examination of residuals shows that these are weak effects.

INSERT Table 4 about here

INSERT Table 5 about here

INSERT Table 2 about here

4.1 Metrics

INSERT Table 6 about here

We now apply the metrics suggested earlier. We take as an example direct and indirect routes from $A =$ **Aff** to $C =$**Nat**. Table 6 shows the counts and calculations for $A \Rightarrow C$; in all there are $n \times n = 81$ such tables to construct for this data set. A visualization of the flows for this pair is shown in Figure 3. The top node is the source node $A =$ **Aff**. 10.3% of all journeys begin with this node, which is drawn with area proportional to 10.3% as a visual clue to its importance as a starting node. The destination node, $C =$ **Nat**, is drawn with area proportional to 7.8%, reflecting the volume of clicks for this node. Shown are the direct and indirect routes. The area of central nodes is not meaningful, these are simple labels. The widths of lines connecting nodes shows how much traffic is flowing between them. The thickest width is between $A =$ **Aff** and $B =$ **Aff**, representing 2749 clicks from A to B which then proceed to another node. The text at the bottom gives the proportion of journeys reaching the destination directly and indirectly. We see that most of the routes from **Aff** to **Nat** are direct, with smaller contributions via **Aff**, **Nat**, **NatB**, and **PPCB**. 145 of the three-step transitions from **Aff** via **Aff** went on to **Nat**, and these 145 clicks represented 14.74% of the direct and indirect transitions from **Aff** to **Nat**. To avoid cluttering the graphic, we avoid drawing flow from intermediary routes if it is less than 2.5% (as an arbitrary threshold) of the number of all routes from **Aff** to **Nat**. One immediate conclusion is that although there are many routes from **Aff** to an intermediary, few of these then continue to **Nat**.

INSERT Figure 3 about here

INSERT Figure 4 about here

Figure 4 concerns the same two nodes, but reversed so that we are exploring routes from **Nat** to **Aff**. For these routes, the great majority are direct and a large proportion of the remainder are via $B =$ **Aff**. We can try to explore several such graphs in parallel; however the task becomes daunting as we need to explore n^2 graphs in all.

Figure 5 summarises the more important journeys via intermediary nodes. For each (source,destination) pair, a stars plot is shown. This shows the proportion of journeys via each kind of intermediate node. To avoid clutter, we show only intermediary nodes accounting for at least 10% of the journey, and bear in mind that we do not show directly the proportion of direct transitions, which can be inferred by the absence of segments showing indirect transitions. The colour and angle of segments is the same for each intermediary. From such a plot we may discern a number of features, depending on the particular example. Here, for example, we note that **PPC** appears to be an important intermediary for destination **PPC**; **NatB** is an important intermediary whenever the source or destination node is **NatB**; and **Aff** is an important intermediary whenever the source or destination node is **Aff**.

INSERT Figure 5 about here

4.2 Channel relevance

INSERT Table 7 about here

Table 7 shows the relative value of nodes from three perspectives. The first represents the volume of two-step journeys starting at a node. The second represents the volume of two-step journeys ending at a node. The third shows the relative value of a node as an intermediary using the formulae derived to (4). The main interpretation is that **NatB** is very important in the journey, but not quite so much as would be believed simply looking at source and destination information. On the other hand, channel **Aff** has a slightly more important role than source and destination information suggests. Otherwise there are few major differences between channels for this retailer.

5 Discussion

Ultimately, the deep question is whether a specified channel actually matters, and this is the central issue in this paper, in which we address how such relevance can be inferred from data, with a focus on whether a channel is important in a customer journey as a means of transiting from a source to a destination node. We have provided a metric based on three-step transitions in order to measure this importance. Statistical sequential pattern analysis of this kind is highly challenging: one aim of future work is to examine longer journey fragments. A second theme of future work is to explore the roles of intermediary nodes in determining conversion behaviour; however we would need to collect meaningful data about non-converting journeys in order to do this, and this would require being careful about the assumptions of non-converting journeys. We have not taken into account the value of conversion; for example it may be that some nodes are relevant only for low-revenue conversions. A third theme of future work is to provide a usable means of hypothesis testing, taking into account the vast number of potential hypotheses of interest.

Acknowledgements

Part of this research was funded by Knowledge Transfer Partnership KTP007499, funded by Summit Media Ltd. and by the UK Technology Strategy Board. We are grateful to Summit Media Ltd for providing data and to colleagues there for providing expertise.

References

V. Abhishek, P. Fader, and K. Hosanagar. The long road to online conversion: A model of multi-channel attribution. http://dx.doi.org/10.2139/ssrn.2158421, 2012.

Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. Technical report, IBM Research Division, Almaden Research Center, 1995.

Bettina Berendt and Myra Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. *VLDB J.*, 9(1):56–75, 2000.

Sule Gunduz and M. Tamer Ozsu. A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 535–540, 2003.

Sule Gunduz-Oguducu and M. Tamer Ozsu. Incremental click-stream tree model: Learning from new users for web page prediction. *Distributed and Parallel Databases*, 19(1):5–27, 2006.

Michael Hahsler, Bettina Grün, and Kurt Hornik. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, October 2005. ISSN 1548-7660. URL http://www.jstatsoft.org/v14/i15.

Internet Advertising Bureau UK. 2012 Online Adspend Full Year Results. http://www.iabuk.net/research/library/2012-full-year-digital-adspend-results, 2013.

A. Jamalzadeh. Statistical methods for ecommerce. Phd thesis, Durham University, 2012.

Xuhiu Shao and Lexin Li. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 258–264, 2011.

W. Wang and J. Yang. *Mining Sequential Patterns from Large Data Sets*. Springer, New York, 2005.

D. A. Wooff and J. M. Anderson. Time-weighted attribution of revenue to multiple online marketing channels in the customer journey in e-commerce. *Durham Research Online*, 2013.

D. A. Wooff and A. Jamalzadeh. Robust and scale-free effect sizes for non-normal two-sample comparisons, with applications in e-commerce. *Journal of Applied Statistics*, 40(11):2495–2515, 2013.

L. Xu, J. A. Duan, and A. B. Whinston. Path to purchase: A mutually exciting point process model for online advertising and conversion. http://dx.doi.org/10.2139/ssrn.2149920, 2012.

Mohammed J. Zaki. Sequence mining in categorical domains: Algorithms and applications. In Ron Sun and Lee Giles, editors, *Sequence Learning: Paradigms, Algorithms, and Applications*, volume 1828 of *LNAI State-of-the-Art-Survey*, pages 162–187. Springer-Verlag, Heidelberg, Germany, 2000a.

Mohammed J. Zaki. Sequence mining in categorical domains: Incorporating constraints. In *Proceedings of the 9th International Conference on Information and Knowledge Management, Washington D.C*, pages 422–429, 2000b.

Table 1: Single-click-journey probabilities

Channel	Code	Freq	Prob
Affiliates	Aff	3841	0.1401
Banner	Ban	62	0.0023
Price Comparison	Comp	818	0.0298
Listed Referrer	List	96	0.0035
Natural Search (Other)	Nat	1954	0.0713
Natural Search (Brand)	NatB	14081	0.5135
Pay-per-click	PPC	2174	0.0793
Pay-per-click (Brand)	PPCB	2543	0.0927
Unlisted Referrer	Un	1851	0.0675
All		27420	1.0000

Table 2: Bivariate transition counts, $N\{ik\}$

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	4374	53	289	24	579	1892	467	516	400
Ban	52	40	15	2	24	162	37	28	71
Comp	476	25	342	3	194	528	183	144	62
List	40	3	7	35	18	189	20	30	23
Nat	1052	32	208	26	2199	2239	682	483	277
NatB	3172	174	511	144	2023	29320	1586	1924	1385
PPC	939	44	262	25	839	2161	2719	666	288
PPCB	857	45	135	27	434	2148	499	2955	506
Un	567	87	61	13	222	1122	251	344	6387

Table 3: Counts of transitions from sender to receiver via the **NatB** node (B), $X_i \rightarrow$ **NatB** $\rightarrow X_k$.

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	274	8	33	7	94	577	68	57	29
Ban	4	6	2	0	0	83	5	6	13
Comp	38	5	34	0	30	177	23	19	4
List	1	0	1	9	8	92	3	5	7
Nat	102	9	29	8	350	829	113	86	50
NatB	701	88	144	82	678	14549	527	555	410
PPC	100	14	34	4	138	789	346	97	21
PPCB	120	7	23	7	81	723	79	323	47
Un	40	7	5	4	36	394	33	34	174

Table 4: Counts of transitions from sender to receiver via any intermediary node, $\sum_j N\{ijk\}$.

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	2400	44	187	21	405	1249	339	333	231
Ban	30	29	12	1	14	115	30	21	52
Comp	226	17	200	2	134	305	124	83	38
List	18	3	3	25	13	132	15	24	14
Nat	529	23	130	19	1470	1367	512	321	161
NatB	1551	134	314	117	1399	18188	1136	1209	775
PPC	501	36	164	13	584	1326	1742	447	173
PPCB	471	33	90	22	281	1240	362	1828	289
Un	292	60	36	9	146	710	177	238	4626

Table 5: Proportion of indirect transitions to direct transitions, v_{ik}, for each node pair.

	Receiver, k								
Sender, i	Aff	Ban	Comp	List	Nat	NatB	PPC	PPCB	Un
Aff	0.55	0.83	0.65	0.88	0.70	0.66	0.73	0.65	0.58
Ban	0.58	0.72	0.80	0.50	0.58	0.71	0.81	0.75	0.73
Comp	0.47	0.68	0.58	0.67	0.69	0.58	0.68	0.58	0.61
List	0.45	1.00	0.43	0.71	0.72	0.70	0.75	0.80	0.61
Nat	0.50	0.72	0.62	0.73	0.67	0.61	0.75	0.66	0.58
NatB	0.49	0.77	0.61	0.81	0.69	0.62	0.72	0.63	0.56
PPC	0.53	0.82	0.63	0.52	0.70	0.61	0.64	0.67	0.60
PPCB	0.55	0.73	0.67	0.81	0.65	0.58	0.73	0.62	0.57
Un	0.51	0.69	0.59	0.69	0.66	0.63	0.71	0.69	0.72

Table 6: Metric calculations for the relevance of intermediary nodes, for source node $i = \mathbf{Aff}$ and destination node $k = \mathbf{Nat}$.

Node	Transition count				Metric, %		
j	$N\{ij\}$	$N\{jk\}$	$N\{ijk\}$	$\sum_k N\{ijk\}$	Λ_{ijk}	Φ_{ijk}	Ψ_{ijk}
Aff	4374	579	145	2749	14.7	25.0	3.3
Ban	53	24	1	25	0.1	4.2	1.9
Comp	289	194	16	142	1.6	8.2	5.5
List	24	18	1	15	0.1	5.6	4.2
Nat	579	2199	89	332	9.0	4.0	15.4
NatB	1892	2023	94	1147	9.6	4.6	5.0
PPC	467	839	21	264	2.1	2.5	4.5
PPCB	516	434	27	326	2.7	6.2	5.2
Un	400	222	11	209	1.1	5.0	2.8

Table 7: The value of intermediary nodes

	Source	Destination	Intermediary
Aff	0.1031	0.1383	0.1694
Ban	0.0052	0.0060	0.0156
Comp	0.0235	0.0219	0.0349
List	0.0044	0.0036	0.0090
Nat	0.0863	0.0783	0.0861
NatB	0.4826	0.4768	0.4204
PPC	0.0953	0.0773	0.0754
PPCB	0.0912	0.0850	0.0947
Un	0.1086	0.1127	0.0945

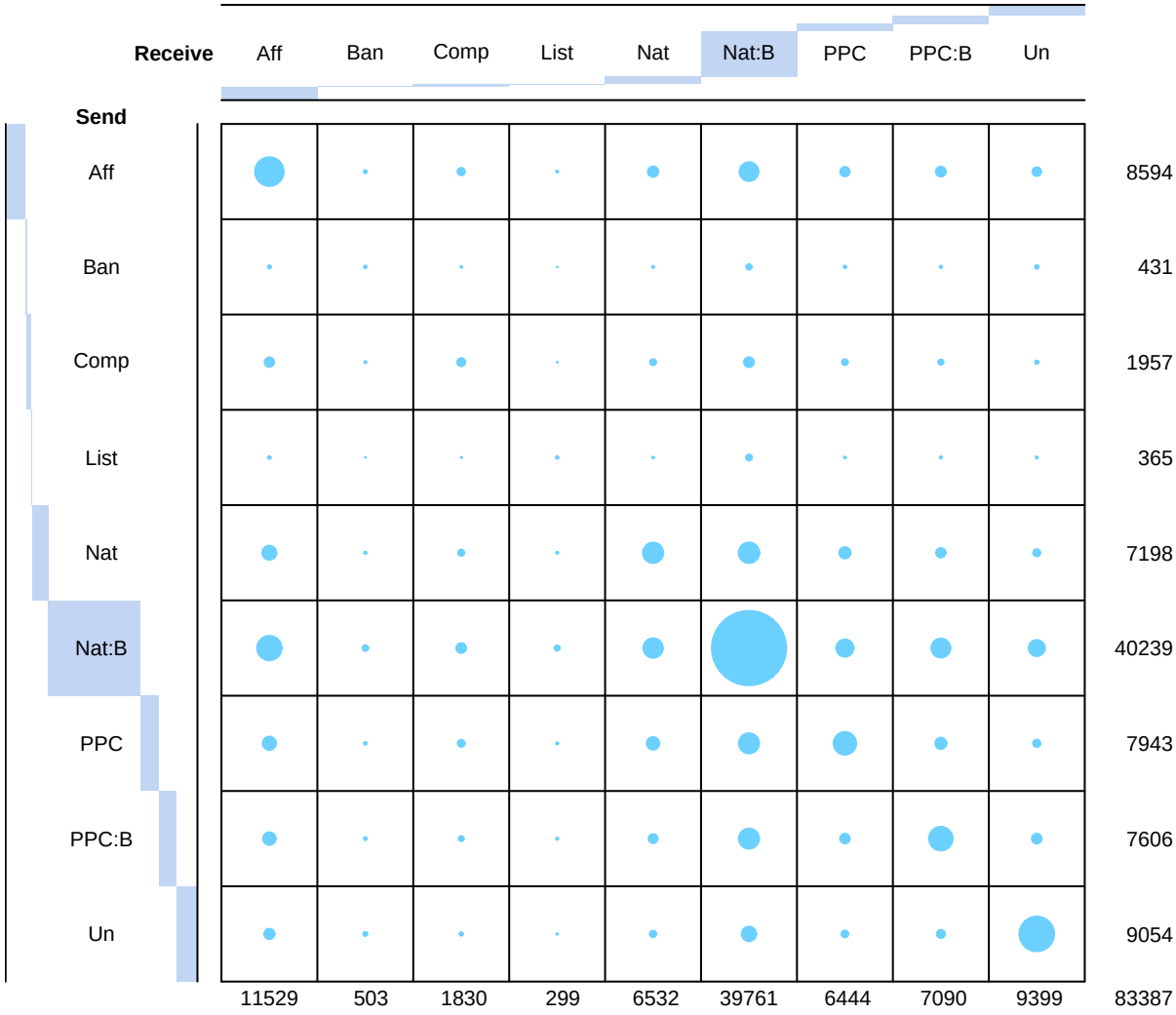

Figure 1: Direct transition frequency as a proportion of all journeys.

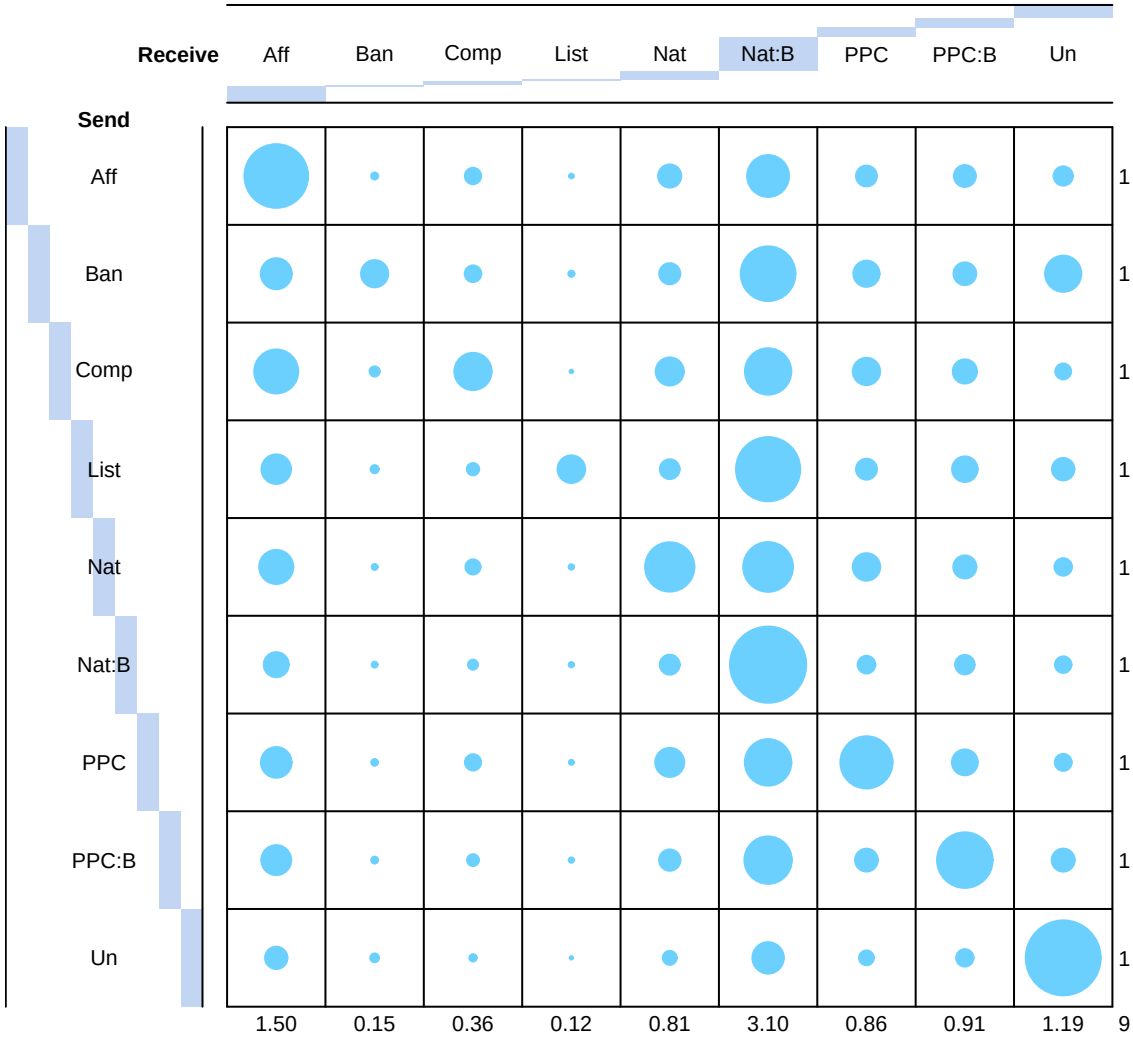

Figure 2: Transition probability given that current state is the sending node.

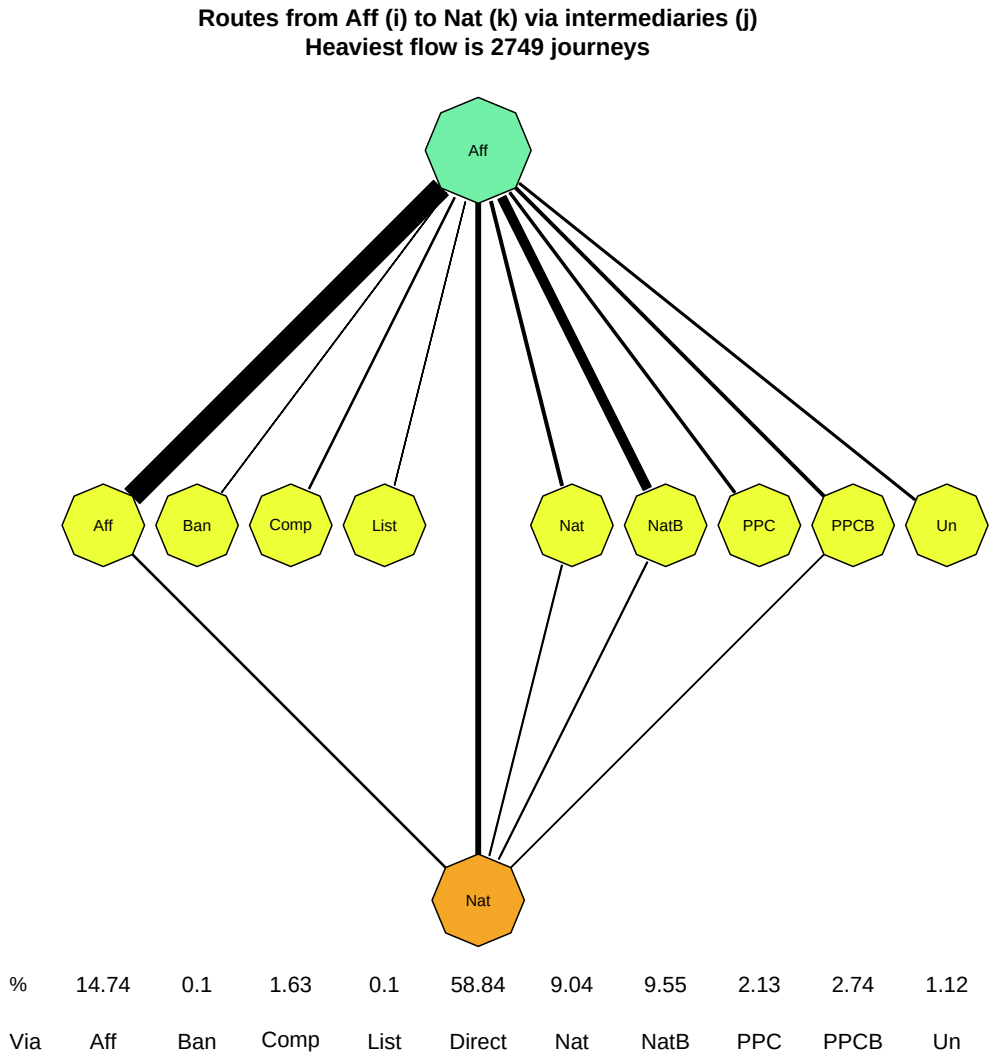

Figure 3: Relevance of intermediate nodes in journeys from **Aff** to **Nat**. Line widths indicate volume. Low volumes are omitted.

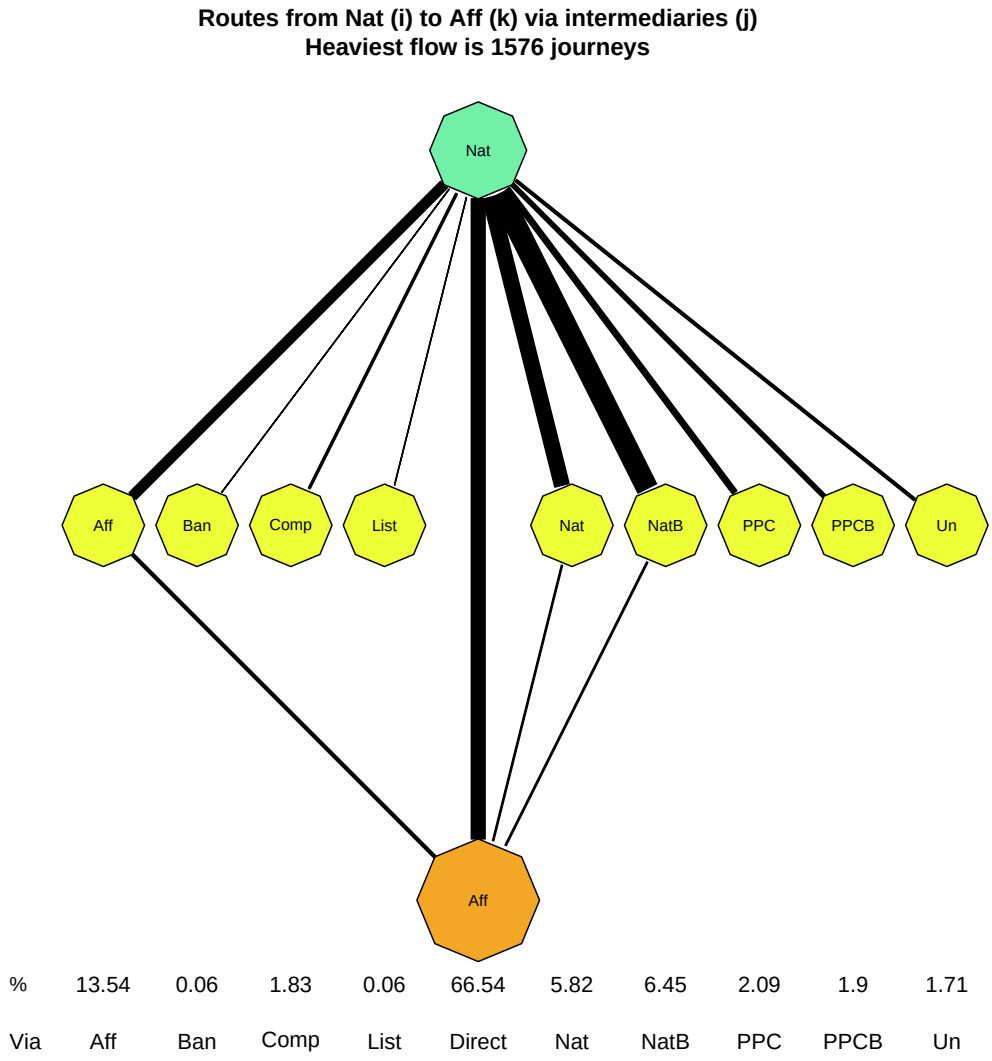

Figure 4: Relevance of intermediate nodes in journeys from **Nat** to **Aff**. Line widths indicate volume. Low volumes are omitted.

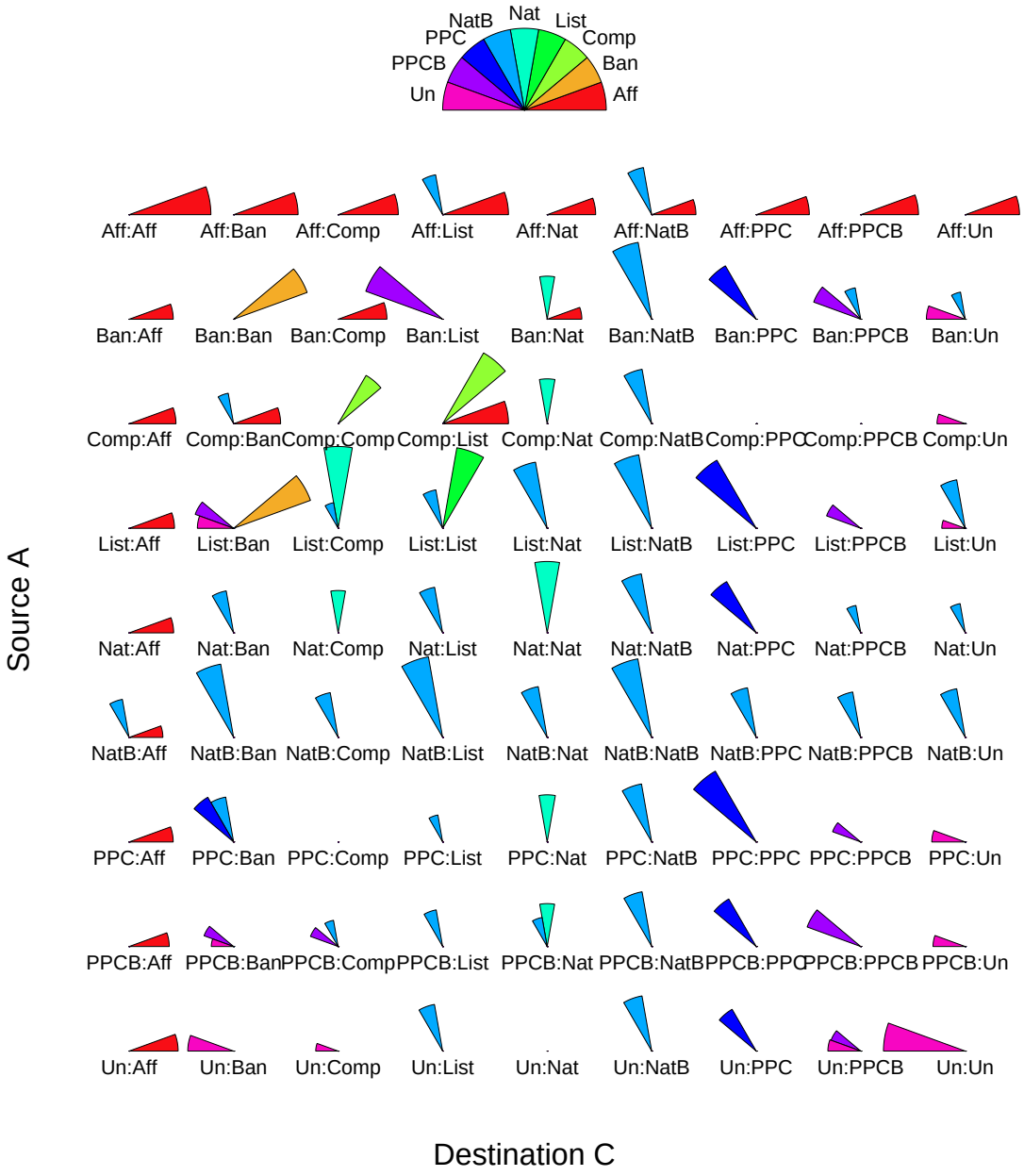

Figure 5: Relevance of intermediate nodes in all journeys. Journeys less than 10% as a proportion are omitted. The colour and angle of segments is the same for each intermediary node B.